Dunning–Kruger Effects in Face Perception

Xingchen Zhou[1], Rob Jenkins[1]*

[1]Department of Psychology, University of York, UK

***Corresponding author**

Rob Jenkins,

Department of Psychology

University of York

York YO10 5DD

UK

rob.jenkins@york.ac.uk

+44 1904 323144

Abstract

The Dunning–Kruger Effect refers to a common failure of metacognitive insight in which people who are incompetent in a given domain are unaware of their incompetence. This effect has been found in a wide range of tasks, raising the question of whether there is any 'special' domain in which it is not found. One plausible candidate is face perception, which has sometimes been thought to be 'special'. To test this possibility, we assessed participants' insight into their own face perception abilities (self-estimates) and those of other people (peer estimates). We found classic Dunning–Kruger Effects in matching tasks for unfamiliar identity, familiar identity, gaze direction, and emotional expression. Low performers overestimated themselves, and high performers underestimated themselves. Interestingly, participants' self-estimates were more stable across tasks than their actual performance. In addition, peer estimates revealed a consistent egocentric bias. High performers attributed higher accuracy to other people than did low performers. We conclude that metacognitive insight into face perception abilities is limited and subject to systematic biases. Our findings urge caution when interpreting self-report measures of face perception ability. They also reveal a fundamental source of uncertainty in social interactions.

Key Words

metacognition; face perception; Dunning-Kruger effect; egocentric bias

1 Introduction

Negotiating everyday life requires that our plans are commensurate with our abilities. This basic requirement underscores the importance of metacognition—insight into one's own thinking and the thinking of others (Fleming, Dolan, & Frith, 2012; Jost, Kruglanski, & Nelson, 1998; Tullis & Fraundorf, 2017). In fact, metacognitive insight is not only inaccurate, it is also subject to systematic biases. One influential example of such a bias is the Dunning–Kruger Effect, encapsulated in the title of its debut paper, "unskilled and unaware of it" (Kruger & Dunning, 1999). The headline result is that poor performers in a given task drastically overestimate their ability, believing that they are outperforming the majority when, in fact, they are the ones being outperformed (Dunning, Johnson, Ehrlinger, & Kruger, 2003). Kruger & Dunning's (1999) explanation of this effect is elegant, and points to a cruel trap in human psychology: The skills that grant competence in a particular domain are the very skills needed to evaluate competence in that domain. People who lack the former lack the latter. A secondary result concerns the top of the ability range. High performers often *underestimate* their standing, but for an entirely different reason. These individuals recognise that they perform well, they just assume that other people perform well too.

Part of the appeal of the Dunning–Kruger Effect is its broad generality. The same basic pattern emerges in reasoning (Kruger & Dunning, 1999; Pennycook, Ross, Koehler, & Fugelsang, 2017), humour (Kruger & Dunning, 1999), political knowledge (Anson, 2018; Motta, Callaghan, & Sylvester, 2018), and many other domains. Indeed, the apparent ubiquity of Dunning–Kruger

Effects has prompted some to wonder if there is any 'special' domain in which the standard pattern is not found (Dunning, Johnson, Ehrlinger, & Kruger, 2003).

One plausible candidate for such a 'special' domain is face perception. Evidence that faces may be cognitively special comes from at least four sources (McKone & Robbins, 2011). First, developmental studies have suggested that newborns demonstrate some innate knowledge of facial structure (Goren, Sarty, & Wu, 1975; Johnson, 2005). Second, face perception seems to be disproportionately affected by image manipulations such as inversion (Yin, 1969; Rossion, 2008) and contrast reversal (Kemp, Pike, White, & Musselman, 1996; Farroni et al., 2005). Third, it has been proposed that face perception may be subserved by face-specific neural circuitry (Riddoch et al., 2008; Kanwisher & Yovel, 2006). More recently, genetic studies have shown that human face recognition ability is specific and heritable (Wilmer et al., 2010; Wilmer 2017). This converging evidence from highly diverse studies has led some researchers to propose that face perception may involve specialised or face-specific cognitive processes.

Despite the theoretical and applied interest in face processing, no previous studies have tested for Dunning–Kruger Effects in this domain. A few studies have found that individuals in the general population show minimal to moderate insight into their own face recognition abilities (e.g. Bindemann, Attard, & Johnston, 2014; Palermo et al., 2017; Bobak, Mileva, & Hancock, 2019), echoing findings for other types of memory (Beaudoin & Desrichard, 2011; but see Livingston & Shah, 2018; Arizpe et al., 2019 for more positive views). However, none of these studies was concerned with metacognition in the 'expansive' sense that includes insight into

other people's abilities (Jost, Kruglanski, & Nelson, 1998). Their main interest was whether a person's self-report (e.g. agreement with questionnaire items such as, "My face recognition ability is worse than most people"; Shah et al., 2015) could predict the same person's performance on standard face recognition tests. They did not compare estimated performance and actual performance for the same task.

A few face perception studies have examined other aspects of metacognition. Sauerland et al (2016) adapted the choice blindness paradigm (Johansson, Hall, Sikström, & Olsson, 2005) to investigate insight into identification judgements. Participants were asked to sort photographs of unfamiliar faces by identity (Jenkins, White, Van Montfort, & Burton, 2011). They were then confronted with one of their identity decisions and asked to justify it. On critical trials, the photographs were secretly switched, so that the decisions participants were asked to justify were opposite to the decisions that they actually made. Very few of these manipulations were detected. Indeed, participants readily reported their reasoning behind identity decisions that they had not reached.

Such findings suggest that insight into one's own face recognition performance is somewhat limited. Fewer studies have examined insight into other people's face recognition performance. Ritchie et al. (2015) presented pairs of faces in a matching task for identity. As expected, participants performed better with familiar faces than with unfamiliar faces (Clutterbuck & Johnston, 2004; Noyes & Jenkins, 2017, 2019). However, participants also predicted that the faces they themselves knew would be easier for other people to match—even people who did

not know those faces. These findings demonstrate an egocentric bias in identification performance (Greenwald, 1980), in that viewers estimated the cognition of others from their own perspective (DiMaggio et al., 2008; Hinds, 1999; Kelley & Jacoby, 1996). However, it remains unclear whether high-performing participants produced higher estimates than low-performing participants.

Given that face perception is often presented as a special case for cognition, we tested whether it is a special case for metacognition. Specifically, we asked whether standard Dunning–Kruger Effects and egocentric bias emerge in face perception tasks. We begin in Experiment 1 with identification tasks for familiar and unfamiliar faces. In Experiment 2, we expand our analysis to include other aspects of face perception, namely gaze direction, and emotional expression.

2 Experiment 1. Identity matching for familiar and unfamiliar faces

Our first experiment had two main aims. First, we sought to establish whether face perception follows the same metacognitive principles as other aspects of cognition. Specifically, we asked whether Dunning–Kruger Effects and egocentric bias are observed in face identification tasks. Second, we sought to compare these metacognitive patterns for familiar and unfamiliar faces. To address these questions, we adapted a standard perceptual matching task for facial identity (Burton, White, McNeill, 2010). In the standard task, participants are presented with pairs of face photos. For each pair, the task is to decide whether the two photos show the same person (50% of trials) or different people (50% of trials). Accuracy on this task is typically at ceiling for familiar faces (e.g. Clutterbuck & Johnston, 2004; Noyes & Jenkins, 2017, 2019), but is generally

much lower for unfamiliar faces (e.g. Clutterbuck & Johnston, 2004; Burton, White, McNeill, 2010; Noyes & Jenkins, 2017, 2019).

This task has several characteristics that make it well suited to comparison of cognition and metacognition. First, each trial has a correct answer, so accuracy can be scored objectively. Second, the same/different response options mean that ceiling performance and chance performance are well defined (100% accuracy and 50% accuracy respectively). Third, there are large individual differences in performance (White, Kemp, Jenkins, Matheson, & Burton, 2014), such that high- and low-scoring respondents tend to be clearly separated. Recording actual scores allows us to assign participants to performance quartiles, as per Dunning & Kruger (1999). Recording participants' estimated scores allows us to test (i) whether 'incompetent' participants (lowest performance quartile) show the classic 'unskilled and unaware' pattern, and (ii) whether 'competent' participants (highest performance quartile) underestimated their performance.

Previous studies of metacognition have often relied on retrospective estimates of performance, collected after the whole task (e.g. Dunning & Kruger, 1999; Tenenberg & Murphy, 2005; Simons, 2013; Feld, Sauermann & de Grip, 2017; see Sarac & Karakelle, 2012; Gignac & Zajenkowski, 2020, for useful discussions of this issue). That approach has several drawbacks. One is that it imposes substantial demands on retrospective memory. Cognitive tasks often involve dozens of trials or items, and these will typically vary in subjective difficulty. The challenge is not only to recall the landscape of that experience, but also to encapsulate it in a single score. To complicate matters, the overall impression may be skewed by primacy and

recency effects (Haugtvedt & Wegener, 1994). To avoid these issues, we captured (i) actual performance, (ii) self-estimates, and (iii) peer estimates on each trial. We also captured participants' self-estimates of their own percentile ranking at the end of each task, for 'backward compatibility' with previous studies.

In light of Dunning & Kruger's (1999) findings, we predicted that low performers would overestimate their performance and that high performers would underestimate their performance. Since virtually everyone is a high performer for familiar face identification (Burton, Wilson, Cowan, & Bruce, 1999; Jenkins & Kerr, 2013), we expected this interaction to be compressed (near ceiling) for familiar faces. In light of the egocentric bias (Ross, Greene, & House, 1977), we expected low performers to make low peer estimates, and high performers to make high peer estimates. It follows that peer estimates should be lower for unfamiliar faces than for familiar faces.

2.1 Method

2.1.1 Participants

Sixty-four UK students (44 female, 20 male; mean age 20 years; age range 18–26 years) from the University of York took part in exchange for a small payment or course credit. The experiments in this study were approved by the Ethics Committee at the University of York. All participants provided written informed consent.

2.1.2 Stimuli and apparatus

Ambient images of 20 familiar faces (e.g. UK and US celebrities; 10 female, 10 male) and 20 unfamiliar faces (e.g. celebrities from other countries; 10 female, 10 male) were downloaded from online sources. Each image was cropped and resized to 570 pixels high × 380 pixels wide for onscreen presentation. For Different Person trials, we paired faces that resembled each other and matched the same basic verbal description (e.g. young woman with red hair). To avoid image repetition, we collected four photos of each face—two for use in Same Person trials, and two for use in Different Person trials. Each face appeared in Same and Different trials equally often, and each participant saw each image exactly once. To ensure that all participants received identical tasks, all participants received identical image pairings. Experiments were run using a 21.5-inch iMac with i5 processor. Stimulus presentation and data collection were controlled by PsychoPy2 v1.82.00 (Peirce, 2007, 2008).

2.1.3 Design

All participants completed both the *Familiar* and the *Unfamiliar* face matching task in separate blocks. Block order was counterbalanced so that half of the participants encountered the *Familiar* condition first, and half of them encountered the *Unfamiliar* condition first. Within each block, the 40 trials (20 *Same* person, 20 *Different* person) were presented in a random order. All participants contributed the same measures in both tasks—actual performance, self-estimates, and peer estimates.

We defined *actual performance* as actual test score, that is, the proportion of correct responses in the matching task. Participants' actual test scores were used to determine their actual

percentile ranking (0–100%), and to define performance quartiles for the Dunning–Kruger analyses.

Self-estimates comprised two metrics. *Estimated test score* was an estimate of absolute performance, captured trial-by-trial. Following each identity decision (Same or Different), participants indicated whether they were sure or unsure of their decision. Estimated test score was defined as the number of 'sure' responses plus half of the number of 'unsure' responses. That is, we assumed that participants guessed on unsure trials and answered half of them correctly by chance. *Estimated percentile ranking* (0–100%) was an estimate of relative performance, reported by each participant at the end of each task.

Peer estimates were also captured trial-by-trial. For each image pair, participants estimated the proportion of respondents who would answer correctly (0–20%, 21–40%, 41–60%, 61–80%, 81–100%). To provide context for these estimates, participants were informed that all respondents were UK students. We note that estimated percentile ranking, reported at the end of each task, combines self-estimate and peer estimate.

2.1.4 Procedure

Each display consisted of a pair of face photographs alongside a set of response options as shown in Figure 1.

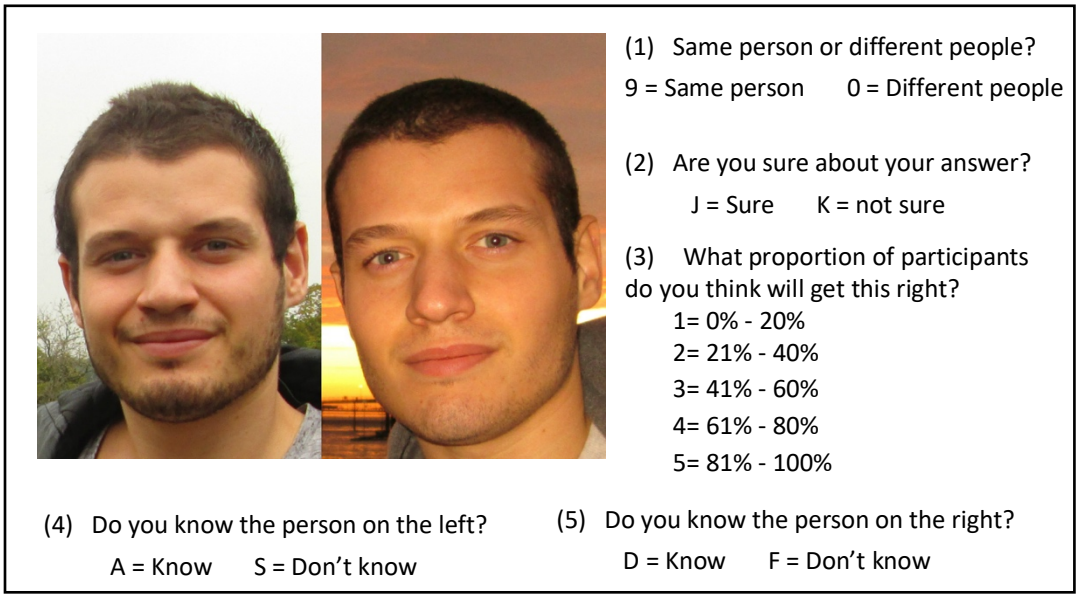


Figure 1. Example identity matching display from Experiment 1. In this example, the two photos show the same unfamiliar face. Participants respond to questions 1–5 for each image pair.

For each pair, participants indicated (i) whether the two photos showed the *Same* person or *Different* people, (ii) whether they were *Sure* or *Unsure* of their decision, (iii) the proportion of participants they thought would give the correct answer, and (iv) whether or not they knew the face in each image. Participants were reminded that they did not need to know the person's name to know that person's face. Each display remained on screen until the final response, which immediately initiated the next trial. The experimenter explained the task at the beginning of the session using a printed example display, which showed a face that was not presented in the main experiment. Following this example trial, each participant underwent two blocks of 40 trials each (one Familiar block and one Unfamiliar block). At the end of each block, participants estimated their own performance relative to all participants (percentile ranking) by dragging an

onscreen slider (0%, “I think I performed worse than other participants” to 100%, “I think I performed better than other participants”). Participants were able to rest between blocks and initiated the next block by pressing the space bar. The entire test session took approximately 30 minutes to complete.

2.2 Results and discussion

Faces in the Familiar condition were familiar to more participants (M = 55, SE = 1.05) than faces in the Unfamiliar condition (M = 5, SE = .43) [t (158) = 44.51, p < .001, d = 7.04], confirming that our familiarity manipulation was successful.

In comparing cognition and metacognition, we first examined participants’ insight into their own absolute performance (test score) and relative performance (percentile ranking), by combining actual attainment with self-estimates in the same analyses. Our main focus is the Dunning–Kruger analysis based on performance quartiles. We then examined participants’ insight into other people’s performance, focusing specifically on egocentric bias.

2.2.1 Insight into one’s own performance

Dunning and colleagues (Dunning et al., 2003) established the convention of analysing metacognition data by performance quartiles. In this approach, participants are divided into quartiles according to their actual performance. Estimated performance can then be compared to actual performance in each quartile. Figure 2 summarizes this analysis for test score and percentile ranking, separately for the Familiar and Unfamiliar face matching tasks.

Test scores

Familiar face matching

Participants' test scores were submitted to a 2 × 4 mixed ANOVA with the within-subjects factor of Measure (Actual Score, Estimated Score) and the between-subjects factor of Quartile (Lowest, Second, Third, Highest). This analysis revealed a main effect of Measure, with Estimated scores (M = 96.61, SE = .49) exceeding Actual scores (M= 95.43, SE = .35) overall [F (1,60) = 5.62, p < .05, η_p^2 = .09]. Unsurprisingly, there was also a main effect of Quartile, with scores increasing from the lowest quartile to the highest quartile [F (3, 60) = 37.76, p < .001, η_p^2 = .65]. In keeping with the standard Dunning–Kruger pattern, these main effects were qualified by a significant Measure × Quartile interaction [F (3,60) = 9.58, p < .001, η_p^2 = .32]. Simple main effects showed that Estimated score exceeded Actual score in the Lowest quartile [F (1,60) = 26.74, p < .001, η_p^2 = .31], but not in the 2nd [F (1,60) = 1.81, p = .18], 3rd [F (1,60) = .09, p = .76] or Highest quartiles [F (1,60) = 3.28, p = .08]. The simple main effect of Quartile was significant for both Actual scores [F (3,60) = 68.35, p < .001, η_p^2 = .77] and Estimated scores [F (3,60) = 7.58, p < .001, η_p^2 = .28].

Unfamiliar face matching

Test scores for the unfamiliar face matching task were analysed in the same way. For unfamiliar faces, there was no difference between Estimated scores (M = 79.99, SE = 1.28) and Actual scores (M = 82.05, SE = .39) overall [F (1,60) = 2.83, p = .10]. Again, there was a main effect of Quartile, with scores increasing from the lowest quartile to the highest quartile [F (3, 60) = 15.69, p < .001, η_p^2 = .44]. There was also a significant crossover interaction between these

factors [F (3,60) = 8.42, p < .001, η_p^2 = .30]. Simple main effects showed that Estimated score exceeded Actual score in the Lowest quartile [F (1, 60) = 5.28, p < .05, η_p^2 = .08] but not the 2nd quartile [F (1,60) = .78, p = .38]. The effect then reversed in the 3rd [F (1, 60) = 13.02, p < .01, η_p^2 = .18] and highest quartiles [F (1, 60) = 9.73, p < .01, η_p^2 = .14], such that Actual score exceeded Estimated score. The simple main effect of Quartile was significant for Actual scores [F (3, 60) = 133.57, p < .001, η_p^2 = .87], but not for Estimated scores [F (3,60) = 1.53, p = .22].

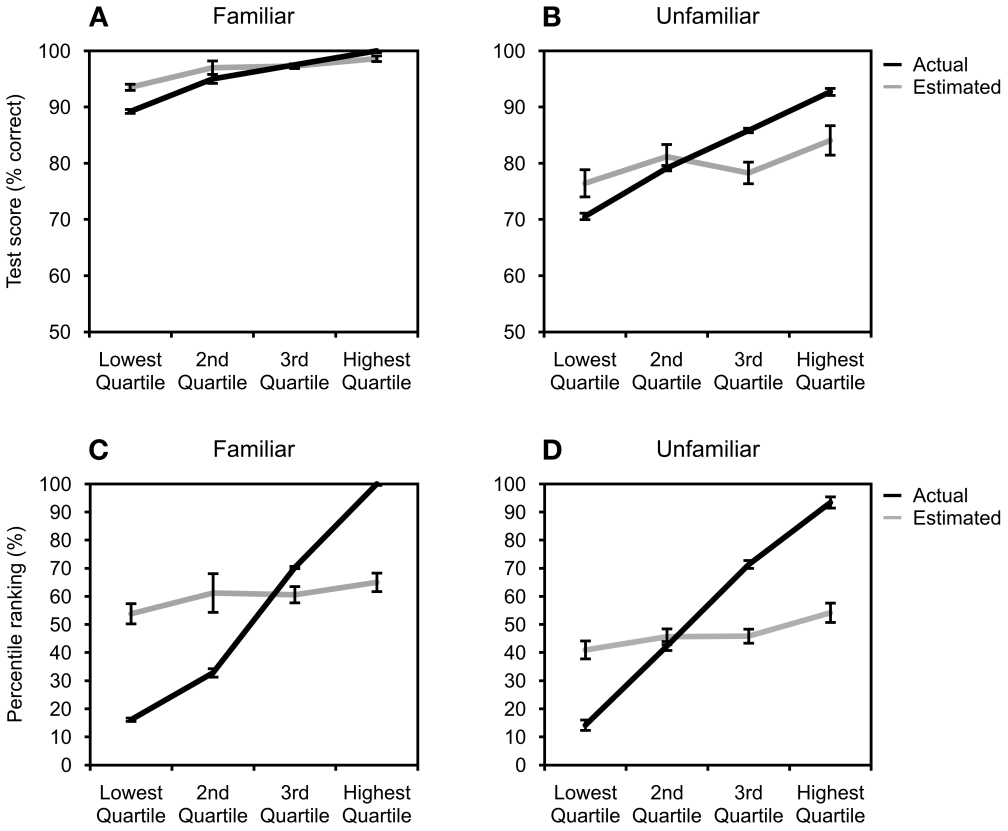


Figure 2. Dunning–Kruger analysis of the face matching tasks in Experiment 1. The top row shows test scores for (A) Familiar faces and (B) Unfamiliar faces. Actual scores (black) and Estimated scores (grey) are plotted separately for each performance quartile. Chance performance is 50%. The bottom row shows percentile rankings for (C) Familiar faces and (D) Unfamiliar faces. Actual ranks (black) and Estimated ranks (grey) are plotted separately for each performance quartile. Error bars show SE.

Percentile ranking

Familiar face matching

As with the test scores, percentile rankings were entered into a 2 × 4 mixed ANOVA with the within-subjects factor of Measure (Actual Score, Estimated Score) and the between-subjects factor of Quartile (Lowest, Second, Third, Highest). This analysis revealed a main effect of Measure, with Estimated rank (M = 60.15, SE = 2.46) exceeding Actual rank (M= 54.81, SE = .66) overall [F (1, 60) = 4.65, p < .05, η_p^2 = .07], and the expected main effect of Quartile [F (3, 60) = 91.83, p < .001, η_p^2 = .82]. There was also a significant crossover interaction between Measure and Quartile [F (3, 60) = 62.82, p < .001, η_p^2 = .76]. Simple main effects showed that Estimated rank exceeded Actual rank in the Lowest quartile [F (1, 60) = 82.47, p < .001, η_p^2 = .58] and the 2nd quartile [F (1, 60) = 14.63, p < .001, η_p^2 = .20]. However, this effect was reversed in the 3rd [F (1, 60) = 8.22, p < .01, η_p^2 = .12] and Highest quartiles [F (1, 60) = 84.46, p < .001, η_p^2 = .59], in which Actual rank exceeded Estimated rank. The simple main effect of Quartile was significant for Actual rank [F (3, 60) = 1137.10, p < .001, η_p^2 = .98], but not for Estimated rank [F (3,60) = 1.35, p = .27].

Unfamiliar face matching

Percentile ranks for the unfamiliar face task were analysed in the same way. This analysis revealed a significant effect of Measure, with Actual rank (M = 55.31, SE = 1.11) exceeding Estimated rank (M= 46.65, SE = 1.75) [F (1, 60) = 17.16, p < .001, η_p^2 = .22], and the expected main effect of Quartile, with ranks increasing from the lowest quartile to the highest quartile [F (3, 60) = 82.98, p < .001, η_p^2 = .81]. As with the familiar face task, there was a significant

crossover interaction between Measure and Quartile [F (3, 60) = 46.16, p < .001, η_p^2 = .70]. Simple main effects showed that Estimated rank exceeded Actual rank in the Lowest quartile [F (1, 60) = 37.38, p < .001, η_p^2 = .38] but not the 2nd quartile [F (1,60) = .68, p = .41]. This effect was reversed in the 3rd [F (1, 60) = 50.81, p < .001, η_p^2 = .46] and Highest quartiles [F (1, 60) = 68.90, p < .001, η_p^2 = .54], with Actual rank exceeding Estimated rank. The simple main effect of Quartile was significant for Actual rank [F (3, 60) = 219.02, p < .001, η_p^2 = .92], but not for Estimated rank [F (3,60) = 2.06, p = .12].

2.2.2 Insight into other people's performance

To assess egocentric bias in each task, we compared peer estimates (attributions of other people's performance) generated by the highest and lowest performing participants. Egocentric bias predicts that peer estimates from the Highest quartile will be higher than peer estimates from the Lowest quartile. Figure 3 summarises this analysis separately for the familiar and unfamiliar face matching tasks.

Familiar face matching

For each familiar face, we calculated the mean peer estimate from Lowest quartile and Highest quartile participants. Peer estimates were on a scale of 1–5, where 1 means "0–20% of participants will answer correctly", and 5 means "81–100% of participants will answer correctly" (see Figure 1). An independent t-test confirmed that peer estimates from Highest quartile participants (M = 4.67, SE = .03) were significantly higher than those from Lowest quartile participants (M = 4.37, SE = .06) [t (78) = 4.72, p < .001, d = 1.06].

Unfamiliar face matching

Peer estimates in the unfamiliar face matching task were analysed in the same way. An independent t-test showed that peer estimates were higher for the Highest quartile (*M* = 3.81, *SE* = .05) than for the Lowest quartile (*M* = 3.56, *SE* = .05) [*t* (78) = 3.70, *p* < .001, *d* = .83].

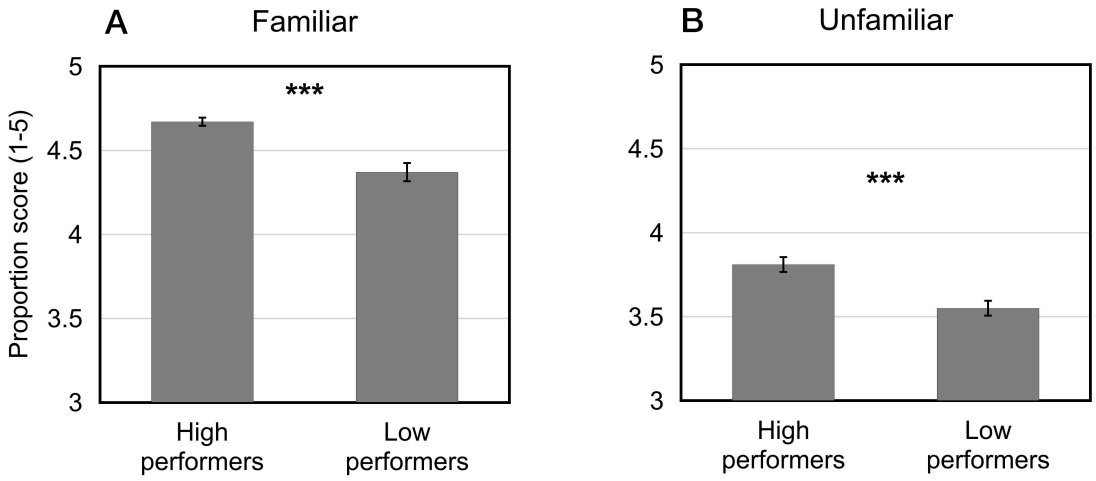


Figure 3. Egocentric bias in peer estimates from the face matching tasks in Experiment 1. (A) Familiar faces. (B) Unfamiliar faces. In both tasks, High performers attributed higher performance to others; Low performers attributed lower performance to others. Error bars show SE.

The Dunning–Kruger analysis of test scores (Figure 2) showed that self-estimates were higher for familiar faces than for unfamiliar faces. Combining this observation with egocentric bias implies that peer estimates should also be higher for familiar faces than for unfamiliar faces. A within-subjects t-test confirmed that this difference was significant (Familiar *M* = 4.59, *SE* = .04; Unfamiliar *M* = 3.69, *SE* = .04) [*t* (63) = 19.27, *p* < .001, *d* = 2.41].

The cognitive aspects of these results were as expected from previous research. When matching faces for identity, accuracy was at ceiling for familiar faces (95% correct overall), and was significantly lower for unfamiliar faces (82% correct overall). We also obtained the expected individual differences in performance. Although unfamiliar face matching was generally poor, some people were much better at it than others (range 60–97.5%). This wide range in performance lends itself to a Dunning–Kruger type of analysis.

Claims of face being 'special' notwithstanding, we found absolutely standard Dunning–Kruger effects in face identification. Low performers overestimated their performance, and high performers underestimate their performance. This pattern emerged in test score (an absolute measure, captured trial by trial), and in percentile rank (a relative measure, captured retrospectively). It also occurred in both Familiar and Unfamiliar identity conditions, though test scores in the Familiar condition were somewhat compressed against ceiling.

We also saw a clear evidence of egocentric bias. High performers made higher peer estimates than low performers; and peer estimates were higher overall for familiar faces than for unfamiliar faces.

All of these findings concern matters of identification. Given that other aspects of face perception (such as gaze direction and emotional expression) are known to dissociate from identification, we next examined metacognition for these other tasks.

3 Experiment 2. Identity, gaze, and expression matching

The purpose of our second experiment was to establish whether the metacognitive pattern seen for identification in Experiment 1 extends to other face tasks. Specifically, we asked whether Dunning–Kruger Effects and egocentric bias extend to perception of gaze direction and emotional expression. These tasks are especially interesting from a metacognition perspective. First, gaze direction and emotional expression are dissociable from face identification (Andrews & Ewbank, 2004; Hoffman & Haxby, 2000; Winston, Henson, Fine-Goulden, & Dolan, 2004). This dissociation allows us to test the generalizability of metacognitive patterns across cognitively unrelated tasks. Second, unlike perception of facial identity, perception of gaze direction and emotional expression have both been associated with cognitive insight, in the specific sense of inferring other people's mental states from their behaviour (Calder et al., 2002; Friesen & Kingstone, 1998; Simpson & Crandall, 1972). Given that the ability to infer mental states seems related to metacognition, it is possible that individuals who perform especially well in these tasks will also demonstrate especially high metacognitive insight (and vice versa).

To extend our analysis to 'cognitive insight' signals from the face, we adapted the identity matching task from Experiment 1 to assess perception of gaze direction and emotional expression. To allow replication of key findings, and to facilitate comparison across diverse tasks, we also repeated the unfamiliar face matching task from Experiment 1. The task format (Same/Different judgements to paired images) and task measures (Actual versus Estimated test scores and percentile ranks) were the same in all three tasks. This homology ensured that data from all three tasks could be analysed in the same way.

Based on previous studies, we expected actual performance on the identity, gaze, and expression tasks to be either uncorrelated (identity versus gaze; identity versus expression) or weakly correlated (gaze versus expression). Our main interest was whether similar metacognitive patterns emerged in all three tasks. If our gaze and expression tasks require metacognitive insight, then people with the greatest insight should perform best, and people with the least insight should perform worst. In that case, the Dunning–Kruger Effect and the egocentric bias should break down. On the other hand, if metacognitive biases generalize even across tasks that are not correlated at the cognitive level, then the Dunning–Kruger Effect and the egocentric bias should persist in all three tasks.

3.1 Method

3.1.1 Participants

Sixty-four UK students (56 female, 8 male; mean age = 20 years; age range 18–26 years) from the University of York took part in exchange for a small payment or course credit. None of these volunteers participated in Experiment 1.

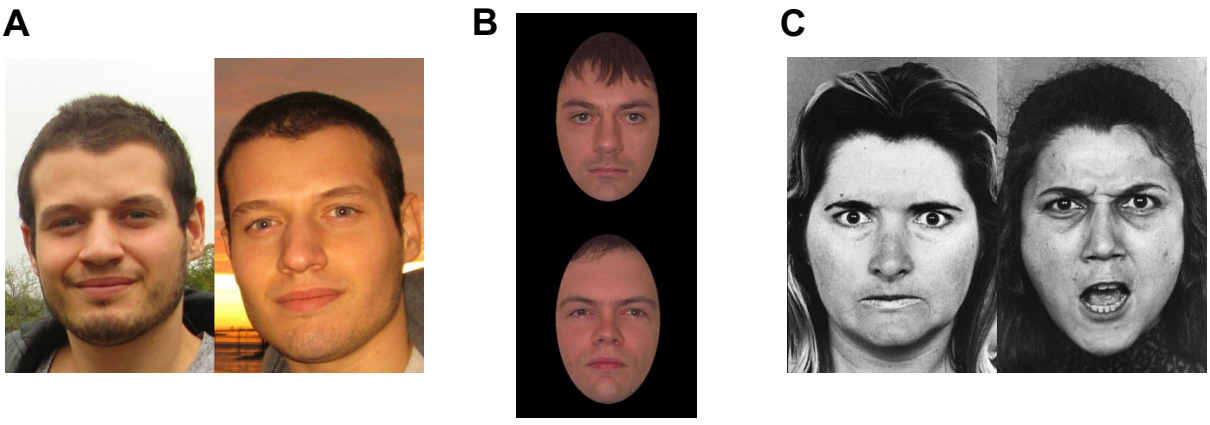
3.1.2 Stimuli

Face identity task

Stimuli for the identity matching task were the same as for the unfamiliar face matching task in Experiment 1 (See Figure 1 and Figure 4). As all of the faces were now unfamiliar, we omitted the image-by-image familiarity check (Questions 4 & 5 in Figure 1).

Gaze direction task

Stimuli for the gaze matching task were drawn from Jenkins, Beaver, and Calder (2006). We selected eight models (4 female, 4 male), each posing five gaze directions (10° left [L10], 5° left [L05], straight ahead [S00], 5° right [R05], 10° right [R10]; 40 images in total). Each face was presented in an elliptical mask measuring 230 pixels high × 205 pixels wide. Stimulus pairs always combined two identities of the same sex. For each combination, we created a Same Direction pair (two faces looking in the same direction: L10, L05, S00, R05, or R10) and a Different Direction pair (two faces looking in different directions). To ensure a range of difficulty, Different Direction pairs differed by 5° (S00 vs R05; S00 vs R05), 10° (L05 vs R05), or 20° (L10 vs R10; R10 vs L10). To make deviations from the midline easier to discern, the two faces in each pair were arranged vertically rather than horizontally (see Figure 4). Each face appeared once at



the top and once at the bottom in both a Same Direction and a Different Direction trial, resulting in a total of 80 trials.

Figure 4. Example face matching stimuli from Experiment 2. (A) Identity matching. (B) Gaze matching. (C) Expression matching.

Expression task

Stimuli for the expression matching task were drawn from the Facial Expressions of Emotion: Stimuli and Tests (FEEST) dataset (Young, Perrett, Calder, Sprengelmeyer, & Ekman, 2002). Given that facial expressions of happiness are reliably recognized (Ekman, Friesen, & Ellsworth, 1972; Calvo & Lundqvist, 2008), we excluded happiness images to avoid ceiling effects. We selected five female models, each posing five facial expressions of emotion (anger, disgust, fear, surprise, and sadness; 25 images in total). Each face image measured 362 pixels high × 241 pixels wide. Stimulus pairs always combined two identities. Each image was combined with each identity in a Same Emotion pair (two faces expressing the same emotion) and a Different Emotion pair (two faces expressing different emotions), resulting in a total of 100 trials. The two images in each pair were arranged horizontally (see Figure 4). Each identity and each emotion appeared equally often on the left and on the right.

3.1.3 Design

All participants completed the *Identity*, *Gaze*, and *Expression* matching tasks in separate blocks. Block order was counterbalanced with respect to participants so that each task could be encountered first, second, or third. Within each block, trials were presented in a random order. All participants contributed the same measures in all three tasks—actual performance, self-estimates, and peer estimates.

3.1.4 Procedure

The procedure was the same as in Experiment 1 except for the following changes. Participants now completed three matching tasks (*Identity*, *Gaze Direction*, *Emotional Expression*), making *Same*/*Different* judgements according to the task. As before, participants indicated whether they were *Sure* or *Unsure* of each decision, and estimated the proportion of participants (UK students) they thought would give the correct answer. The entire test session took approximately 40 minutes to complete.

3.2 Results and discussion

Before proceeding to the metacognitive analyses, we first examined performance on each of the three face matching tasks. At the group level, actual scores were very similar for the three tasks (Identity M= 78.91, SE = .39; Gaze M= 80.54, SE = .41; Expression M= 82.21, SE = .30), indicating similar levels of overall difficulty. Importantly however, there was no significant correlation between actual scores in the Identity and Gaze tasks [r (62) = .13, p = .31], or between the Identity and Expression tasks [r (62) = .15, p = .25]. There was a moderate correlation between actual scores in the Gaze and Expression tasks [r (62) = .30, p < .05]. For actual rankings, there were no significant correlations between any of the tasks [Identity and Gaze r (62) = .15, p = .25; Identity and Expression r (62) = .20, p = .12; Gaze and Expression r (62) = .17, p = .17]. In sum, the pattern of performance is as expected based on previous work. Invariant and changeable aspects of faces cleave together to some extent, but correlations between different face tasks are otherwise low.

Our metacognitive analysis follows the same plan as Experiment 1. We first examine participants' insight into their own absolute performance (test score) and relative performance (percentile ranking), by combining actual attainment with self-estimates in a Dunning–Kruger analysis for each task. We then examine participants' insight into other people's performance, focusing specifically on egocentric bias. Finally, we consider the stability of cognition and metacognition across different face tasks.

3.2.1 Insight into one's own performance

As in Experiment 1, participants were divided into quartiles according to actual performance. Estimated performance was then compared to actual performance in each quartile. Figure 5 summarises this analysis for test score and percentile ranking, separately for the Identity, Gaze, and Expression matching tasks.

Test scores

Identity matching

Test scores were submitted to a 2 × 4 mixed ANOVA with the within-subjects factor of Measure (Actual Score, Estimated Score) and the between-subjects factor of Quartile (Lowest, Second, Third, Highest). This analysis revealed a main effect of Measure, with Estimated scores (M = 83.03, SE = 1.04) exceeding Actual scores (M= 78.91, SE = .39) overall [F (1, 60) = 14.25, p < .001, η_p^2 = .19]. There was also a main effect of Quartile, with scores increasing from the lowest quartile to the highest quartile [F (3, 60) = 29.28, p < .001, η_p^2 = .59]. These main effects were qualified by a significant Measure × Quartile interaction [F (3, 60) = 19.46, p < .001, η_p^2 = .49].

Simple main effects showed that Estimated score exceeded Actual score in the Lowest quartile [F (1,60) = 59.69, p < .001, η_p^2 = .50] and the 2nd quartile [F (1,60) = 6.73, p < .05, η_p^2 = .10], but not for the 3rd quartile [F (1,60) = .12, p = .73]. The effect then reversed in the Highest quartiles [F (1,60) = 8.14, p < .01, η_p^2 = .12]. The simple main effect of Quartile was significant for Actual scores [F (3,60) = 192.92, p < .001, η_p^2 = .91] but not for Estimated scores [F (3,60) = .61, p = .61].

Gaze matching

Test scores for the gaze matching task were analysed in the same way. Again, there was a main effect of Measure, with Estimated scores (M = 91.59, SE = .76) exceeding Actual scores (M= 80.54, SE = .41) overall [F (1, 60) = 166.87, p < .001, η_p^2 = .74] and a main effect of Quartile, with scores increasing from the lowest quartile to the highest quartile [F (3, 60) = 20.03, p < .001, η_p^2 = .50]. These main effects were also qualified by a significant Measure × Quartile interaction [F (3, 60) = 17.48, p < .001, η_p^2 = .47]. Simple main effects showed that Estimated score exceeded Actual score in the Lowest quartile [F (1,60) = 145.31, p < .001, η_p^2 = .71], the 2nd quartile [F (1,60) = 64.21, p < .001, η_p^2 = .52] and the 3rd quartile [F (1,60) = 20.64, p < .001, η_p^2 = .26], but not for the Highest quartiles [F (1,60) = 3.21, p = .08]. The simple main effect of Quartile was significant for Actual scores [F (3,60) = 83.37, p < .001, η_p^2 = .81] but not for Estimated scores [F (3,60) = .56, p = .65].

Expression matching

For the Expression task, there was a main effect of Measure, with Estimated scores (M = 89.69, SE = .83) exceeding Actual scores (M= 82.21, SE = .30) overall [F (1,60) = 75.33, p < .001, η_p^2 = .56]

and a main effect of Quartile, with scores increasing from the lowest quartile to the highest quartile [F (3,60) = 9.67, p < .001, η_p^2 = .33]. These main effects were qualified by a significant Measure × Quartile interaction [F (3, 60) = 16.97, p < .001, η_p^2 = .46]. Simple main effects showed that Estimated score exceeded Actual score in the Lowest quartile [F (1,60) = 94.51, p < .001, η_p^2 = .61] and the 2nd quartile [F (1,60) = 37.12, p < .001, η_p^2 = .38], but not for the 3rd quartile [F (1,60) = 2.62, p = .11] or the Highest quartiles [F (1,60) = .17, p = .68]. The simple main effect of Quartile was significant for Actual scores [F (3,60) = 112.10, p < .001, η_p^2 = .85] but not for Estimated scores [F (3,60) = .50, p = .69].

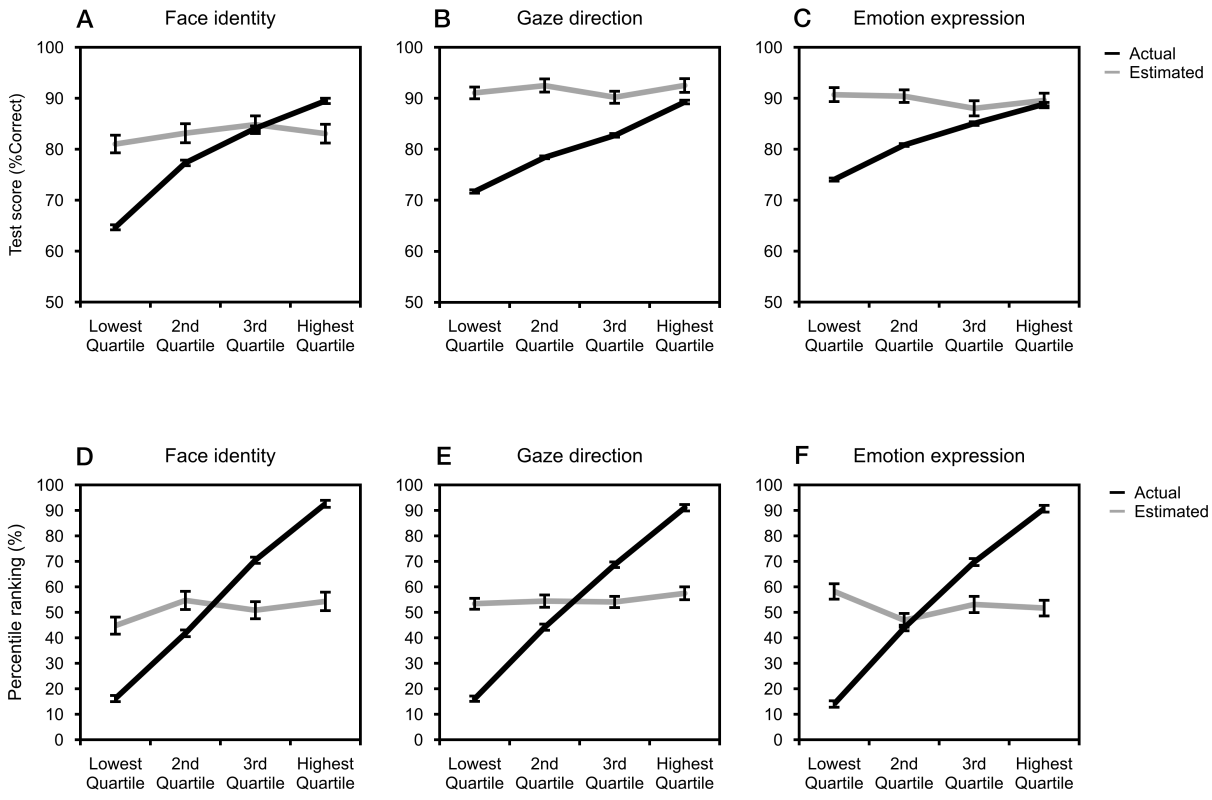


Figure 5. Dunning–Kruger analysis of the face matching tasks in Experiment 2. The top row shows test scores for (A) Identity, (B) Gaze, and (C) Expression. Actual scores (black) and Estimated scores (grey) are plotted separately for each performance quartile. Chance performance is 50%. The bottom row shows

percentile rankings for (D) Identity, (E) Gaze, and (F) Expression. Actual ranks (black) and Estimated ranks (grey) are plotted separately for each performance quartile. Error bars show SE.

Percentile ranking

Identity matching

As with the test scores, percentile rankings were entered into a 2 × 4 mixed ANOVA with the within-subjects factor of Measure (Actual Score, Estimated Score) and the between-subjects factor of Quartile (Lowest, Second, Third, Highest). On this occasion, the overall difference between Estimated rank (M = 51.19, SE = 2.01) and Actual rank (M = 55.29, SE = .91) was not significant [F (1, 60) = 3.58, p = .06]. There was the expected main effect of Quartile [F (3, 60) = 67.01, p < .001, η_p^2 = .77] and a significant crossover interaction between Measure and Quartile [F (3, 60) = 49.41, p < .001, η_p^2 = .71]. Simple main effects showed that Estimated rank exceeded Actual rank in the Lowest quartile [F (1, 60) = 46.70, p < .001, η_p^2 = .44] and the 2nd quartile [F (1, 60) = 8.44, p < .01, η_p^2 = .12]. However, this effect was reversed in the 3rd [F (1, 60) = 22.01, p < .001, η_p^2 = .27] and Highest quartiles [F (1, 60) = 73.78, p < .001, η_p^2 = .55], in which Actual rank exceeded Estimated rank. The simple main effect of Quartile was significant for Actual rank [F (3, 60) = 334.93, p < .001, η_p^2 = .94], but not for Estimated rank [F (3,60) = 1.33, p = .27].

Gaze matching

Again, there was no overall difference between Estimated rank (M = 54.88, SE = 1.44) and Actual rank (M = 55.05, SE = .85) [F (1, 60) = .01, p = .91]. There was a main effect of Quartile [F (3, 60) = 85.36, p < .001, η_p^2 = .81] and a significant crossover interaction between Measure and Quartile

[F (3, 60) = 109.75, p < .001, η_p^2 = .85]. Simple main effects showed that Estimated rank exceeded Actual rank in the Lowest quartile [F (1, 60) = 179.62, p < .001, η_p^2 = .75] and the 2nd quartile [F (1, 60) = 11.34, p < .01, η_p^2 = .16]. However, this effect was reversed in the 3rd [F (1, 60) = 26.15, p < .001, η_p^2 = .30] and Highest quartiles [F (1, 60) = 113.40, p < .001, η_p^2 = .65], in which Actual rank exceeded Estimated rank. The simple main effect of Quartile was significant for Actual rank [F (3, 60) = 364.52, p < .001, η_p^2 = .95], but not for Estimated rank [F (3,60) = .37, p = .78].

Expression matching

As with the other tasks, there was no overall difference between Estimated rank (M = 52.52, SE = 1.78) and Actual rank (M = 54.62, SE = .92) [F (1, 60) = 1.19, p = .28]. Again, the results showed the expected main effect of Quartile [F (3, 60) = 57.28, p < .001, η_p^2 = .74] and a significant crossover interaction between Measure and Quartile [F (3, 60) = 82.56, p < .001, η_p^2 = .81]. Simple main effects showed that Estimated rank exceeded Actual rank in the Lowest quartile [F (1, 60) = 133.19, p < .001, η_p^2 = .69] but not for the 2nd quartile [F (1, 60) = .75, p = .39]. This effect was reversed in the 3rd [F (1, 60) = 16.50, p < .001, η_p^2 = .22] and Highest quartiles [F (1, 60) = 97.43, p < .001, η_p^2 = .62], in which Actual rank exceeded Estimated rank. The simple main effect of Quartile was significant for Actual rank [F (3, 60) = 322.44, p < .001, η_p^2 = .94], but not for Estimated rank [F (3,60) = 1.88, p = .14].

3.2.2 Insight into other people's performance

Peer estimates in the three tasks were analysed in the same way as in Experiment 1. Figure 6 summarises the results of this analysis.

Identity matching

Despite the numerical difference, peer estimates from Highest quartile participants (M = 4.13, SE = .05) were not significantly higher than those from Lowest quartile participants (M = 4.03, SE = .03) [t (78) = 1.82, p = .07, d = .41].

Gaze matching

As expected, peer estimates from Highest quartile participants (M = 4.45, SE = .04) were significantly higher than those from Lowest quartile participants (M = 4.23, SE = .03) [t (158) = 4.42, p < .001, d = .70].

Expression matching

Here too, peer estimates from Highest quartile participants (M = 4.46, SE = .03) were significantly higher than those from Lowest quartile participants (M = 4.28, SE = .02) [t (198) = 5.00, p < .001, d = .71].

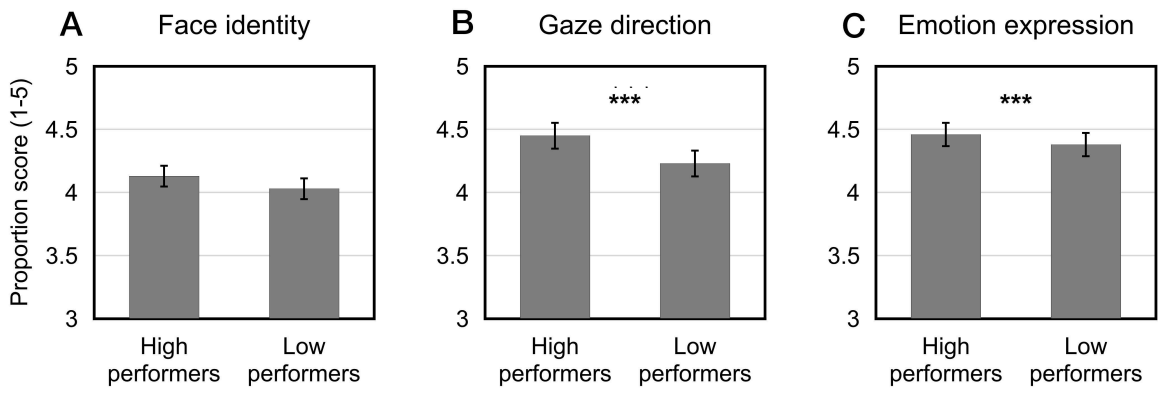


Figure 6. Egocentric bias in peer estimates from the (A) Identity, (B) Gaze, and (C) Expression matching tasks in Experiment 2. High performers attributed higher performance to others; Low performers attributed lower performance to others. Error bars show SE.

One interesting aspect of these findings concerns the Dunning–Kruger analysis of test scores (Figure 5). For high performers in the Gaze and Expression tasks, Actual Scores and Estimated scores converged, but did not cross over. On its own, this pattern may appear to support the idea that gaze and expression perception and metacognition have some shared basis: those who performed best on these matching tasks also showed the most accurate insight into their performance. However, two observations caution against this interpretation. First, high performers did not show accurate insight in the relative measure (percentile ranking; Figure 5), or when estimating the performance of others (Figure 6). Second, estimated scores in the Gaze and Expression tasks were high for all performance quartiles. Why should people think they are so good at these particular tasks? One possibility is poor calibration. Everyday life might provide less useful feedback on errors of gaze and expression (which can vary continuously) compared

with errors of identity (which varies discretely). If that is the case, then people should have less insight into their fallibility in gaze and expression tasks. One way to test this possibility is through feedback training. If people receive feedback on their gaze and expression perception, their self-estimates should fall accordingly.

3.2.3 Stability of cognition and metacognition across face tasks

The preceding analyses show that Dunning–Kruger effects arise in a range of different face tasks. In all of these tasks, low performers overestimated their performance. For high performers, this tendency was reversed or eliminated. Multiple measures of performance give us the opportunity to examine the stability of Dunning–Kruger effects across tasks. Do people who overestimate themselves in one task also overestimate themselves in the other tasks? Or is assessment of one's own performance (like performance itself) task dependent? Figure 7 shows the stability of performance across tasks.

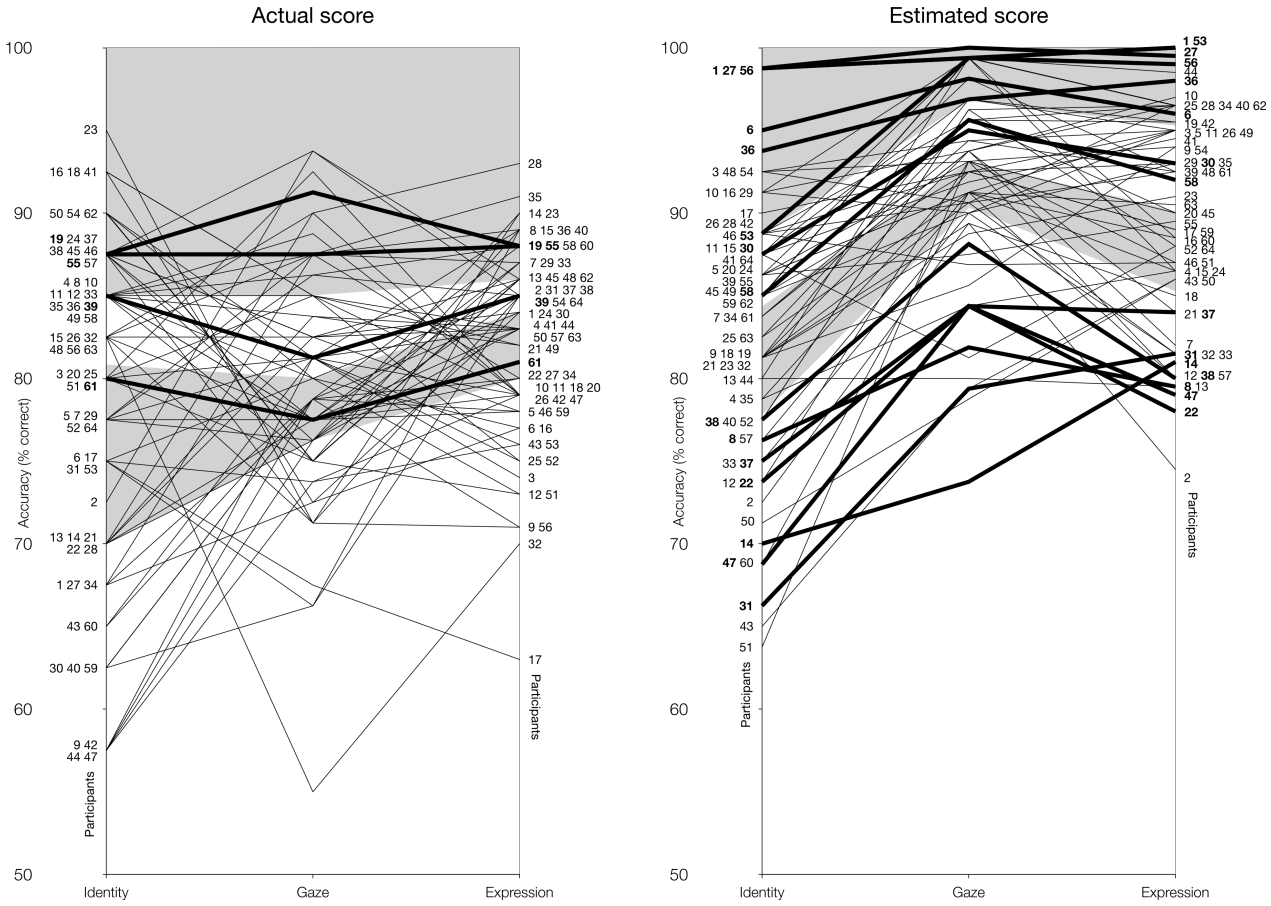


Figure 7. Stability of performance across the three face matching tasks in Experiment 2. Actual scores are shown on the left. Estimated scores are shown on the right. Grey and white regions in each panel are performance quartiles. Heavy lines indicate participants who stayed within the same performance quartile across all three tasks. Light lines indicate participants who switched between performance quartiles.

For Actual Scores, only 4 participants stayed within the same performance quartile across all three tasks. For Estimated Scores, 15 participants stayed within the same quartile. In other words, self-assessment was more stable than ability [χ^2 (1) = 5.26, p < .05; OR = 3.75, 95% CI 1.24–11.30]. This pattern suggests that the tendency to overestimate or underestimate one's

own performance is not strictly task dependent. We return to this issue in the General Discussion section.

4 General Discussion

Unusually for studies of face perception, the experiments reported here concern (i) the cognitive level, (ii) the metacognitive level, and (iii) the relation between these two levels. We first summarize the findings for each of these areas in the context of previous research, before moving on to theoretical and applied implications.

At the cognitive level, performance on the individual face tasks was as expected from previous findings. In the identity matching task, overall accuracy was lower for unfamiliar faces (82% in Experiment 1; 79% in Experiment 2) than for familiar faces (96%), demonstrating the standard familiarity advantage (e.g. Burton, White, & McNeill, 2010; Noyes & Jenkins, 2017, 2019). Paired matching has not been widely used to assess gaze perception or processing of emotional expression, but the present findings demonstrate the applicability of this method to both tasks. Overall accuracy rates were similar across unfamiliar identity, gaze direction, and emotional expression tasks (~80%), and within each task, the range of scores (~55–95%) allowed meaningful analysis of individual differences. Critically, this analysis revealed little or no correlation among scores on the three tasks. That is, a person's score on one task tells us very little about their scores on the other two tasks, even though all three tasks concern face perception. The observed dissociations among these scores are consistent with previous behavioural and neural evidence for independence among face perception abilities (e.g. Young,

Newcombe, de Haan, Small, & Hay, 1993; Duchaine, Jenkins, Germine, & Calder, 2009). However, previous studies have used different tasks, different measures, and different groups to gauge face perception abilities. This is the first time that three such abilities have been assessed in a within-subjects design, using the common task of paired matching. One advantage of this approach is that it imposed the same level for chance performance in all three tasks (50%). This uniformity facilitates comparisons across tasks. It also provides a stable baseline against which to compare metacognitive judgements of one's own and other people's ability.

At the metacognitive level, our findings concern to two processes—*self*-estimates (insight into one's own cognition) and *peer* estimates (insight into other people's cognition). Our self-estimate measures extend Kruger & Dunning's (1999) 'unskilled and unaware' effect into the novel domain of face perception. In identity matching for both familiar faces (Experiment 1) and unfamiliar faces (Experiments 1 and 2), low performers overestimated their own absolute accuracy (percent correct score), and high performers underestimated their own absolute accuracy, giving rise to a classic crossover interaction between estimated test score and actual test score. In matching for gaze direction and for emotional expression (Experiment 2), estimated accuracy levels were higher overall than for the identity tasks. Thus, while low performers again overestimated their own accuracy, for high performers this tendency was merely eliminated rather than being reversed as it was in the identity tasks. For relative accuracy (rank), the picture was clear cut. In all four matching tasks (Experiments 1 and 2), low performers overestimated their rank, and high performers underestimated their rank. These

measures are consistent in showing that participants had rather little insight into their own ability—neither their absolute accuracy level, nor their standing in relation to other people.

Our peer estimate measures also showed a consistent pattern. In identity matching for familiar faces (Experiment 1) and unfamiliar faces (Experiments 1 and 2), high performers attributed higher accuracy to other people than did low performers. Similar performance-contingent estimates emerged in the gaze direction and emotional expression tasks (Experiment 2). One possible interpretation of these performance-contingent effects is that participants estimated other people's ability from their own perspective—that is, with an egocentric bias (Ritchie et al., 2015). On this account, high performers presumed that others can do what they themselves can do, while low performers presumed that others cannot do what they themselves cannot do. The metacognitive picture can be summed up as follows. People estimated their own face perception performance with an "unskilled and unaware" bias, and estimated other people's performance with an egocentric bias.

One interesting aspect of our findings is the consistency of *Estimated* performance across tasks, relative to *Actual* performance across tasks. This pattern suggests that self-estimates are not driven solely by insight into one's own performance, but also involve some determinant that is more stable across tasks. Although the current data do not allow us to single out specific determinants, individual differences in general intelligence or personality could play a role. On a personality account, some participants tend to imagine that they are doing rather well, irrespective of the task, while others tend to imagine that they are doing rather poorly,

irrespective of the task. Several previous studies have reported effects of personality traits on self-estimates in other cognitive domains outside of face perception (e.g. narcissism, Ames & Kammrath, 2004; Big Five, Soh & Jacobs, 2013). Combining personality measures with face perception tasks could help to explain the stability of self-assessments seen here. One interesting question is whether the same personality traits predict self-estimates across domains, or whether any domain-specificity emerges. For example, narcissism might inflate self-assessments generally, whereas extroversion might disproportionately inflate self-assessment of socially relevant abilities, such as face perception. Combined testing should distinguish these possibilities.

As well as their theoretical interest, our findings have implications for face perception in clinical and forensic settings. Several clinical disorders are characterised by specific face perception deficits. In this context, unreliability of self-estimates could influence engagement with clinical services. People with developmental prosopagnosia often have little insight into their own impaired facial identification (Fine, 2012). People with autism spectrum disorders (ASD) may not be aware that they have trouble reading social signals from faces (see Bishop & Seltzer, 2012; Schriber, Robins, & Solomon, 2014, for discussions of self-insight in ASD). If people do not realise that their ability is outside the normal range, they may not seek appropriate help (Yardley, McDermott, Pisarski, Duchaine, & Nakayama, 2008).

Unreliability of peer estimates also has practical implications. There is some evidence that people attribute above-average face recognition ability to individuals with professional training

and experience. For example, participants in Ritchie et al.'s (2015) study predicted that passport officers would outperform students at unfamiliar face matching. In fact, training and experience have no appreciable impact on face recognition ability (Towler et al., 2019; White, Kemp, Jenkins, Matheson, & Burton, 2014). Specialists are generally indistinguishable from university students in terms of task performance (Burton, Wilson, Cowan, & Bruce, 1999; White, Kemp, Jenkins, Matheson, & Burton, 2014). A dissociation between estimated and actual performance of specialists could help to explain the enduring popularity of photo-ID as a means of identifying people, despite evidence of its unreliability (Ritchie et al., 2015).

In future work, it would be interesting to compare estimated and actual performance of automatic face recognition systems on face perception tasks. Although there is a huge literature on automatic face recognition (Ranjan et al., 2018; Phillips et al., 2018), very little is known about human understanding of its accuracy. For now, we show that Dunning–Kruger effects and egocentric bias both arise in face perception. Our findings urge caution when interpreting self-report measures of face perception ability. They also reveal a fundamental source of uncertainty in social interactions.

References

Ames, D. R., & Kammrath, L. K. (2004). Mind-Reading and Metacognition: Narcissism, not Actual Competence, Predicts Self-Estimated Ability. *Journal of Nonverbal Behavior*, *28*(3), 187–209. https://doi.org/10.1023/b:jonb.0000039649.20015.0e

Andrews, T. J., & Ewbank, M. P. (2004). Distinct representations for facial identity and changeable aspects of faces in the human temporal lobe. *NeuroImage*, *23*(3), 905–913. https://doi.org/10.1016/j.neuroimage.2004.07.060

Anson, I. G. (2018). Partisanship, political knowledge, and the Dunning-Kruger effect. *Political Psychology*, *39*(5), 1173–1192. https://doi.org/10.1111/pops.12490

Arizpe, J. M., Saad, E., Douglas, A. O., Germine, L., Wilmer, J. B., & DeGutis, J. M. (2019). Self-reported face recognition is highly valid, but alone is not highly discriminative of prosopagnosia-level performance on objective assessments. *Behavior Research Methods, 51*(3), 1102-1116. https://doi.org/10.3758/s13428-018-01195-w

Beaudoin, M., & Desrichard, O. (2011). Are memory self-efficacy and memory performance related? A meta-analysis. *Psychological Bulletin*, *137*(2), 211–241. https://doi.org/10.1037/a0022106

Bindemann, M., Attard, J., & Johnston, R. A. (2014). Perceived ability and actual recognition accuracy for unfamiliar and famous faces. *Cogent Psychology*, *1*(1), 986903. https://doi.org/10.1080/23311908.2014.986903

Bishop, S. L., & Seltzer, M. M. (2012). Self-reported autism symptoms in adults with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *42*(11), 2354–2363. https://doi.org/10.1007/s10803-012-1483-2

Bobak, A. K., Mileva, V. R., & Hancock, P. J. B. (2019). Facing the facts: Naive participants have only moderate insight into their face recognition and face perception abilities. In *Quarterly Journal of Experimental Psychology* (Vol. 72, Issue 4, pp. 872–881). https://doi.org/10.1177/1747021818776145

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods*, *42*(1), 286–291. https://doi.org/10.3758/brm.42.1.286

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face Recognition in Poor-Quality Video: Evidence from Security Surveillance. *Psychological Science*, *10*(3), 243–248. https://doi.org/10.1111/1467-9280.00144

Calder, A. J., Lawrence, A. D., Keane, J., Scott, S. K., Owen, A. M., Christoffels, I., & Young, A. W. (2002). Reading the mind from eye gaze. *Neuropsychologia*, *40*(8), 1129–1138. https://doi.org/10.1016/s0028-3932(02)00008-8

Calvo, M. G., & Lundqvist, D. (2008). Facial expressions of emotion (KDEF): identification under different display-duration conditions. *Behavior Research Methods*, *40*(1), 109–115. https://doi.org/10.3758/brm.40.1.109

Clutterbuck, R., & Johnston, R. A. (2004). Matching as an index of face familiarity. In *Visual Cognition* (Vol. 11, Issue 7, pp. 857–869). https://doi.org/10.1080/13506280444000021

Dimaggio, G., Lysaker, P. H., Carcione, A., Nicolò, G., & Semerari, A. (2008). Know yourself and you shall know the other… to a certain extent: Multiple paths of influence of self-reflection on mindreading. *Consciousness and Cognition*, *17*(3), 778–789. https://doi.org/10.1016/j.concog.2008.02.005

Duchaine, B., Jenkins, R., Germine, L., & Calder, A. J. (2009). Normal gaze discrimination and

adaptation in seven prosopagnosics. *Neuropsychologia*, *47*(10), 2029–2036. https://doi.org/10.1016/j.neuropsychologia.2009.03.011

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why People Fail to Recognize Their Own Incompetence. *Current Directions in Psychological Science*, *12*(3), 83–87. https://doi.org/10.1111/1467-8721.01235

Ekman, P., Friesen, W. V., & Ellsworth, P. (1972). CHAPTER II - How Do We Determine whether Judgments of Emotion Are Accurate? In P. Ekman, W. V. Friesen, & P. Ellsworth (Eds.), *Emotion in the Human Face* (Vol. 11, pp. 15–19). Pergamon. https://doi.org/10.1016/b978-0-08-016643-8.50009-4

Farroni, T., Johnson, M. H., Menon, E., Zulian, L., Faraguna, D., & Csibra, G. (2005). Newborns' preference for face-relevant stimuli: effects of contrast polarity. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(47), 17245–17250. https://doi.org/10.1073/pnas.0502205102

Feld, J., Sauermann, J., & de Grip, A. (2017). Estimating the relationship between skill and overconfidence. *Journal of Behavioral and Experimental Economics*, *68*, 18–24. https://doi.org/10.1016/j.socec.2017.03.002

Fine, D. R. (2012). A life with prosopagnosia. *Cognitive neuropsychology*, *29*(5-6), 354-359. https://doi.org/10.1080/02643294.2012.736377

Fleming, S. M., Dolan, R. J., & Frith, C. D. (2012). Metacognition: computation, biology and function. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *367*(1594), 1280–1286. https://doi.org/10.1098/rstb.2012.0021

Friesen, C. K., & Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by

nonpredictive gaze. *Psychonomic Bulletin & Review*, *5*(3), 490–495. https://doi.org/10.3758/bf03208827

Gignac, G. E., & Zajenkowski, M. (2020). The Dunning-Kruger effect is (mostly) a statistical artefact: Valid approaches to testing the hypothesis with individual differences data. *Intelligence, 80*, 101449. https://doi.org/10.1016/j.intell.2020.101449

Goren, C. C., Sarty, M., & Wu, P. Y. (1975). Visual following and pattern discrimination of face-like stimuli by newborn infants. *Pediatrics*, *56*(4), 544–549.

Greenwald, A. G. (1980). The totalitarian ego: Fabrication and revision of personal history. *The American Psychologist*, *35*(7), 603. https://doi.org/10.1037/0003-066x.35.7.603

Haugtvedt, C. P., & Wegener, D. T. (1994). Message Order Effects in Persuasion: An Attitude Strength Perspective. *The Journal of Consumer Research*, *21*(1), 205–218. https://doi.org/10.1086/209393

Hinds, P. J. (1999). The curse of expertise: The effects of expertise and debiasing methods on prediction of novice performance. *Journal of Experimental Psychology. Applied*, *5*(2), 205. https://doi.org/10.1037/1076-898x.5.2.205

Hoffman, E. A., & Haxby, J. V. (2000). Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nature Neuroscience*, *3*(1), 80–84. https://doi.org/10.1038/71152

Jenkins, R., Beaver, J. D., & Calder, A. J. (2006). I thought you were looking at me: direction-specific aftereffects in gaze perception. *Psychological Science*, *17*(6), 506–513. https://doi.org/10.1111/j.1467-9280.2006.01736.x

Jenkins, R., & Kerr, C. (2013). Identifiable images of bystanders extracted from corneal

reflections. *PloS One*, *8*(12), e83325. https://doi.org/10.1371/journal.pone.0083325

Jenkins, R., White, D., Van Montfort, X., & Mike Burton, A. (2011). Variability in photos of the same face. *Cognition*, *121*(3), 313–323. https://doi.org/10.1016/j.cognition.2011.08.001

Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, *310*(5745), 116–119. https://doi.org/10.1126/science.1111709

Johnson, M. H. (2005). Subcortical face processing. *Nature Reviews. Neuroscience*, *6*(10), 766–774. https://doi.org/10.1038/nrn1766

Jost, J. T., Kruglanski, A. W., & Nelson, T. O. (1998). Social Metacognition: An Expansionist Review. In *Personality and Social Psychology Review*, *2*(2), 137–154. https://doi.org/10.1207/s15327957pspr0202_6

Kanwisher, N., & Yovel, G. (2006). The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *361*(1476), 2109–2128. https://doi.org/10.1098/rstb.2006.1934

Kelley, C. M., & Jacoby, L. L. (1996). Adult Egocentrism: Subjective Experience versus Analytic Bases for Judgment. *Journal of Memory and Language*, *35*(2), 157–175. https://doi.org/10.1006/jmla.1996.0009

Kemp, R., Pike, G., White, P., & Musselman, A. (1996). Perception and recognition of normal and negative faces: the role of shape from shading and pigmentation cues. *Perception*, *25*(1), 37–52. https://doi.org/10.1068/p250037

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social*

Psychology, *77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Livingston, L. A., & Shah, P. (2018). People with and without prosopagnosia have insight into their face recognition ability. *Quarterly Journal of Experimental Psychology* , *71*(5), 1260–1262. https://doi.org/10.1080/17470218.2017.1310911

McKone, E., & Robbins, R. (2011). Are Faces Special? In Calder, A., Rhodes, G., Johnson, M., & Haxby, J. (Eds.). (2011). *Oxford handbook of face perception*. Oxford University Press. (pp. 149–176). https://doi.org/10.1093/oxfordhb/9780199559053.013.0009

Morton, J., & Johnson, M. H. (1991). CONSPEC and CONLERN: a two-process theory of infant face recognition. *Psychological Review*, *98*(2), 164–181. https://doi.org/10.1037/0033-295x.98.2.164

Motta, M., Callaghan, T., & Sylvester, S. (2018). Knowing less but presuming more: Dunning-Kruger effects and the endorsement of anti-vaccine policy attitudes. *Social Science & Medicine*, *211*, 274–281. https://doi.org/10.1016/j.socscimed.2018.06.032

Noyes, E., & Jenkins, R. (2017). Camera-to-subject distance affects face configuration and perceived identity. *Cognition*, *165*, 97–104. https://doi.org/10.1016/j.cognition.2017.05.012

Noyes, E., & Jenkins, R. (2019). Deliberate disguise in face identification. *Journal of Experimental Psychology. Applied*, *25*(2), 280–290. https://doi.org/10.1037/xap0000213

Palermo, R., Rossion, B., Rhodes, G., Laguesse, R., Tez, T., Hall, B., Albonico, A., Malaspina, M., Daini, R., Irons, J., Al-Janabi, S., Taylor, L. C., Rivolta, D., & McKone, E. (2017). Do people have insight into their face recognition abilities? *Quarterly Journal of Experimental Psychology* , *70*(2), 218–233. https://doi.org/10.1080/17470218.2016.1161058

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1–2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Peirce, J. W. (2008). Generating Stimuli for Neuroscience Using PsychoPy. *Frontiers in Neuroinformatics*, *2*, 10. https://doi.org/10.3389/neuro.11.010.2008

Pennycook, G., Ross, R. M., Koehler, D. J., & Fugelsang, J. A. (2017). Dunning–Kruger effects in reasoning: Theoretical implications of the failure to recognize incompetence. *Psychonomic Bulletin & Review*, *24*(6), 1774–1784. https://doi.org/10.3758/s13423-017-1242-7

Phillips, P. J., Jonathon Phillips, P., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., Cavazos, J. G., Jeckeln, G., Ranjan, R., Sankaranarayanan, S., Chen, J.-C., Castillo, C. D., Chellappa, R., White, D., & O'Toole, A. J. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. In *Proceedings of the National Academy of Sciences* (Vol. 115, Issue 24, pp. 6171–6176). https://doi.org/10.1073/pnas.1721355115

Ranjan, R., Sankaranarayanan, S., Bansal, A., Bodla, N., Chen, J., Patel, V. M., Castillo, C. D., & Chellappa, R. (2018). Deep Learning for Understanding Faces: Machines May Be Just as Good, or Better, than Humans. *IEEE Signal Processing Magazine*, *35*(1), 66–83. https://doi.org/10.1109/msp.2017.2764116

Riddoch, M. J., Johnston, R. A., Martyn Bracewell, R., Boutsen, L., & Humphreys, G. W. (2008). Are faces special? A case of pure prosopagnosia. In *Cognitive Neuropsychology* (Vol. 25, Issue 1, pp. 3–26). https://doi.org/10.1080/02643290801920113

Ritchie, K. L., Smith, F. G., Jenkins, R., Bindemann, M., White, D., & Burton, A. M. (2015). Viewers base estimates of face matching accuracy on their own familiarity: Explaining the photo-ID

paradox. *Cognition*, *141*, 161–169. https://doi.org/10.1016/j.cognition.2015.05.002

Rossion, B. (2008). Picture-plane inversion leads to qualitative changes of face perception. *Acta Psychologica*, *128*(2), 274–289. https://doi.org/10.1016/j.actpsy.2008.02.003

Ross, L., Greene, D., & House, P. (1977). The "false consensus effect": An egocentric bias in social perception and attribution processes. In *Journal of Experimental Social Psychology* (Vol. 13, Issue 3, pp. 279–301). https://doi.org/10.1016/0022-1031(77)90049-x

Saraç, S., & Karakelle, S. (2017). On-line and off-line assessment of metacognition. *International Electronic Journal of Elementary Education*, *4*(2), 301–315. https://files.eric.ed.gov/fulltext/EJ1068610.pdf

Sauerland, M., Sagana, A., Siegmann, K., Heiligers, D., Merckelbach, H., & Jenkins, R. (2016). These two are different. Yes, they're the same: Choice blindness for facial identity. *Consciousness and Cognition*, *40*, 93–104. https://doi.org/10.1016/j.concog.2016.01.003

Schriber, R. A., Robins, R. W., & Solomon, M. (2014). Personality and self-insight in individuals with autism spectrum disorder. *Journal of Personality and Social Psychology*, *106*(1), 112–130. https://doi.org/10.1037/a0034950

Shah, P., Gaule, A., Sowden, S., Bird, G., & Cook, R. (2015). The 20-item prosopagnosia index (PI20): a self-report instrument for identifying developmental prosopagnosia. *Royal Society Open Science, 2*(6), 140343. https://doi.org/10.1098/rsos.140343

Simons, D. J. (2013). Unskilled and optimistic: overconfident predictions despite calibrated knowledge of relative skill. *Psychonomic Bulletin & Review*, *20*(3), 601–607. https://doi.org/10.3758/s13423-013-0379-2

Simpson, W. E., & Crandall, S. J. (1972). The perception of smiles. *Psychonomic Science*, *29*(4),

197–200. https://doi.org/10.3758/bf03332825

Soh, L., & Jacobs, K. E. (2013). The biasing effect of personality on self-estimates of cognitive abilities in males and females. *Personality and Individual Differences*, *55*(2), 141–146. https://doi.org/10.1016/j.paid.2013.02.013

Tenenberg, J., & Murphy, L. (2005). Knowing what I know: An investigation of undergraduate knowledge and self-knowledge of data structures. *Computer Science Education*, *15*(4), 297–315. https://doi.org/10.1080/08993400500307677

Towler, A., Kemp, R. I., Burton, A. M., Dunn, J. D., Wayne, T., Moreton, R., & White, D. (2019). Do professional facial image comparison training courses work? *PloS One*, *14*(2), e0211037. https://doi.org/10.1371/journal.pone.0211037

Tullis, J. G., & Fraundorf, S. H. (2017). Predicting others' memory performance: The accuracy and bases of social metacognition. *Journal of Memory and Language*, *95*, 124–137. https://doi.org/10.1016/j.jml.2017.03.003

White, D., Kemp, R. I., Jenkins, R., Matheson, M., & Burton, A. M. (2014). Passport officers' errors in face matching. *PloS One*, *9*(8), e103510. https://doi.org/10.1371/journal.pone.0103510

Wilmer, J. B., Germine, L., Chabris, C. F., Chatterjee, G., Williams, M., Loken, E., Nakayama, K., & Duchaine, B. (2010). Human face recognition ability is specific and highly heritable. *Proceedings of the National Academy of Sciences*, *107*(11), 5238-5241. https://doi.org/10.1073/pnas.0913053107

Wilmer, J. B. (2017). Individual differences in face recognition: A decade of discovery. *Current Directions in Psychological Science*, *26*(3), 225-230. https://doi.org/10.1177/0963721417710693

Winston, J. S., Henson, R. N. A., Fine-Goulden, M. R., & Dolan, R. J. (2004). fMRI-adaptation reveals dissociable neural representations of identity and expression in face perception. *Journal of Neurophysiology*, *92*(3), 1830–1839. https://doi.org/10.1152/jn.00155.2004

Yardley, L., McDermott, L., Pisarski, S., Duchaine, B., & Nakayama, K. (2008). Psychosocial consequences of developmental prosopagnosia: a problem of recognition. *Journal of Psychosomatic Research*, *65*(5), 445–451. https://doi.org/10.1016/j.jpsychores.2008.03.013

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*(1), 141. https://doi.org/10.1037/h0027474

Young, A. W., Newcombe, F., de Haan, E. H. F., Small, M., & Hay, D. C. (1993). Face perception after brain injury. In *Brain* (Vol. 116, Issue 4, pp. 941–959). https://doi.org/10.1093/brain/116.4.941

Young, A. W., Perrett, D., Calder, A., Sprengelmeyer, R., & Ekman, P. (2002). Facial expressions of emotion: Stimuli and tests (FEEST). *Bury St. Edmunds: Thames Valley Test Company*.